オープンデータを用いた化学特許情報活用へのアプローチ

―特許情報から化学物質の知識を抽出するために―

An approach to Chemical Patent information with Open Data

株式会社富士通研究所

田中　一成

人工知能研究所　ナレッジ活用 PJ、テキストマイニング技術の研究、特許読解支援システムの開発、ナレッジグラフの研究に従事。

tanaka.kazunari@jp.fujitsu.com

株式会社富士通研究所

池田　紀子

R&D マネジメント本部　企画部、技術士（応用理学／総合技術監理部門）、電子・光学デバイス材料の設計および分析、並列処理および分子モデリングの研究、特許読解支援システムの開発に従事。

1　はじめに

特許の三大分野は、特許庁の審査組織から、機械、化学、情報・通信ネットワークである。また、内閣府の第３期科学技術基本計画（H18～22 年度）の分野別推進戦略では、重点推進４分野（ライフサイエンス、情報通信、環境、ナノテクノロジー・材料）があげられた[1]。これら４分野のうち情報通信を除くと、化学分野の特許（以下、化学特許）が関与するので、化学特許調査を支援する研究は重要と考えられる。化学特許では、権利範囲を確定する特許請求範囲（クレーム）が、他分野では見られない方法で発明を特定する場合がある。そのクレームには、新規化合物（2 種類以上の原子が結合した物質）、新規組成物（複数の化合物が混ざったもの）、化合物の製造方法、物性などがある。

特許調査の理想は、データを効果的・効率的に収集・集約、分析し、知識を発見し、新たな価値を創造することである。特許調査の問題は、高度なスキルに多大な時間と労力が必要な点である。そのため、特許調査支援が望まれている。

化学特許を調査する上での課題の１つに、化合物の命名法[2]や構造表記が多様であるため、化合物名を判別できないというものがある。秒単位で増加する化合物に対して、化合物辞書は追従できない。また、一般的な自然言語処理では、化合物名の途中で分断・抽出され、化合物名を正しく認識できない。

化学特許情報には、様々な知識が蓄積されており、正しく化合物名を判別できるようになれば、化合物に関する情報を抽出できる可能性がある。

第２節では、化学特許と化合物データベース（以下、化合物 DB）の情報について述べる。第３節では、知識を抽出するために必要となる化合物名の抽出について述べる。第４節では、化合物が多くの体系名（構造を表す名称）や別称を持つ理由と、オープンデータによる体系名や別称の同一性判定について述べる。

2　化学特許情報と化合物 DB の情報について

化合物の情報としては、以下のようなものがある。

- 化合物名、別称
- 構造
- 物性情報
- 製造プロセス
- 用途

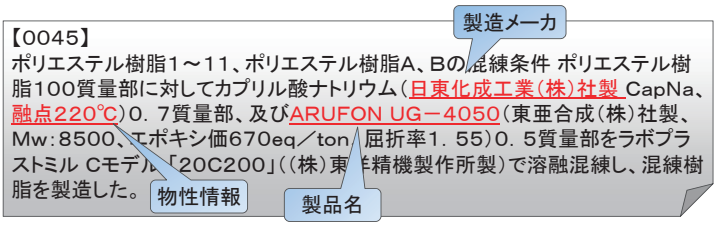

図１　化学特許での化合物の物性情報や製品情報の記述例

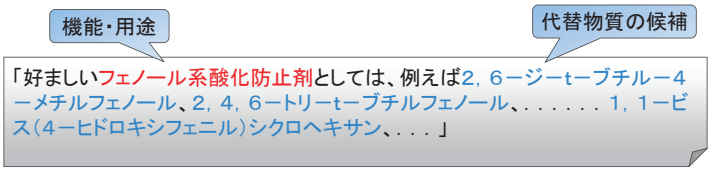

図２　化学特許での機能・用途や代替物質の記述例

・製品情報

これらの情報は、どれも有用であるが、膨大にある化合物についてこれらの情報を収集することは容易ではない。

2.1　化学特許から得られる情報

化合物に関する電子化日本語化学特許は 100 万件を超えており、その中には技術を権利化するための多くの情報が開示されている。

例えば、図 1 に示すように、化合物の組成、融点、分子量などの物性情報や、製造メーカ、製品名、スペックなどの製品情報が書かれる場合がある。また、図 2 のように特許文書中では、機能・用途、代替物質候補群についても書かれる場合が多くある。これらの情報を化学特許全体から抽出することができれば化合物の知識として再利用が可能になる。また、どういう化合物がよく使われているのか、どういう官能基（化合物の部分構造で、物質の化学的属性や化学反応性に着目した概念）が良く出現しているのかを知ることができ、新たな実験への知見や、環境への影響についての情報も得ることができる。

特許情報では、構造式（化合物の構造を原子と原子のつながりによって表したもの）などは画像データとして含んでいるが、画像データは意味付けが容易ではなく、再利用が難しい。

2.2　オープンな化合物 DB から得られる情報と活用への課題

化学特許から抽出する以外にも、これまでに整備されている化合物 DB を利用することで知識を得ることが可能である。代表的な化合物 DB を表 1 に示す[3][4][5][6][7]。

表１　代表的な化合物 DB

データベース	化学物質数	収録開始	ライセンス
日化辞	3,590,000	2005	無料(CC BY 2.1 JP)
ChEMBL	2,100,000	1994	無料(CC BY-SA 3.0)
CAS	132,440,000	1907	有料
PubChem	91,750,000	2004	無料(Public domain)
ChemSpider	59,000,000	2007	無料(CC BY-SA 3.0)

CAS は化学物質の収録数が最大である[5]が、有料であり他のシステムとの連携が難しいため、無料の DB に着目する。

日本化学物質辞書（日化辞）[3]は、有機化合物約 359 万件を収録している。日化辞には化合物の日本語と英語の体系名や別称、構造式や分子式などの情報を収録している。

また、PubChem には、低分子の生化学活性が登録されている[6]。アメリカ合衆国環境保護庁から発行されている CPCat には、化学物質の使用方法と機能や、製品のカテゴリ別リストが登録されている[8]。

このように、化学特許と化合物 DB では、収録されているデータが異なるので、相補的な活用が望まれる。これらの情報をつなげて有効に活用するための課題は、まず、化合物名を抽出すること、また、同じ化合物を表す化合物名の同一性判定をすることである。

3 化合物名の抽出について

特許や論文といった技術文書から、化合物の情報を抽出しようとした場合、化合物名を抽出することが基本になる。

自然言語処理の分野では固有名詞を抽出する技術が知られている。しかし、化合物名の場合には一般的な固有名詞と異なるため、単純に上手くいくとは限らない。日本語の化合物名の特徴はカタカナ、特定の記号、特定の漢字で構成されるという点である。この特徴に当てはまる文字列を抽出することである程度妥当な化合物名の候補を抽出することが可能である。さらに、得られた化合物名の候補について機械学習による分類器を用いることでゴミを除き化合物名を高精度で抽出することができる。

今回、化合物名の抽出手法として、特徴パターン抽出と機械学習を組み合わせる方法を採用した。まず、部分構造名の文字種の組合せパターン（片仮名、英数、「酸」などの漢字、括弧などが連続して並ぶ）から化合物名の候補を抽出する。次に、機械学習を用いて化合物名としてふさわしいかどうかを判別する。

特徴パターン抽出によって得られた化合物名候補7383件について化学の専門家複数人によって、これらの抽出データを正例（例：ビスフェノールA）か負例（例:サービスA）かを仕訳した。その結果、正例2,947件、負例4,436件が得られた。これらの学習データを用いて機械学習により分類器を作成し、化合物名を判別したところ、化合物名の認識精度は、8割程度であった。他のツールによる判別結果と比べて、抽出精度が1.5倍に向上した。

4 化合物名の同一性判定について

化合物の情報を抽出する際に問題なのが、化合物名の同一性判定である。化合物名についてはある程度抽出が可能になってきたが、化合物には多くの体系名や別称があり、どの化合物名が同一の化合物であるのかを判定することは大変困難である。

4.1 化合物名の問題

化合物は以下のような原因で様々な体系名や別称を持つ。

・表記ゆれ
・表記方法の違い
・置換基名の違い
・命名法の違い

例えば、日化辞によると、「フタル酸ジブチル」という化合物では、表2のような体系名と別称を持っている。

体系名や別称の違いは、図3に示す例のように命名規則の違いによって発生する場合もあれば、同じ構造に対して複数の別称が付けられている場合もある。「フタル酸」は、「ベンゼン-1,2-ジカルボン酸」のIUPACで許容されている慣用名（体系的な名称ではないが、広くその分野で用いられている名称）である。

化合物の同一性判定ができない場合には、以下のような問題が発生すると考えられる。

・情報が分散して有効に活用できない

表2　体系名（命名規則の違い）と別称の例

種類	化合物名	備考
体系名	フタル酸ジブチル	官能種類命名法
	ジブチルフタラート	官能種類命名法英語名のカタカナ表記
	ジブチル=フタラート	「ジブチルフタラート」に二重結合記号も表記
	1,2-ベンゼンジカルボン酸ジブチル	置換基命名法(CAS)
	ベンゼン-1,2-ジカルボン酸ジブチル	置換基命名法(IUPAC)
別称	フタル酸n-ブチル	
	n-ブチルフタラート	
	DBP	略称
	エラオル	
	ヘキサプラスM/B	
	Dibutyl phthalate	英語名
	・・・	

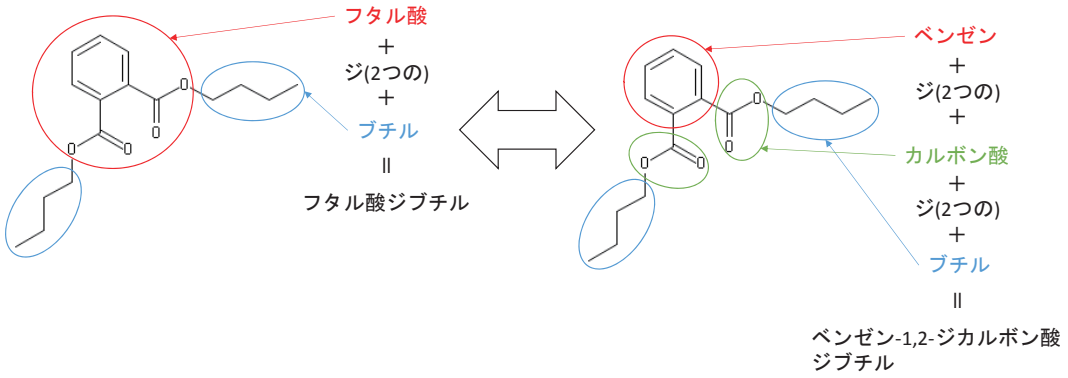

図３　命名規則の違いによる体系名の例

・法規制などを正しく判別できない
・化合物名や官能基の出現回数を集計する場合に正しくカウントできない

集計ができるようになると、特許中の官能基の出現頻度、メーカ、メーカが関与する官能基の出現頻度を知ることができる。Linked Open Data（以下、LOD）と連携すると、メーカの製品情報（官能基の供給元、官能基の消費先など）を示すことができる。

4.2　化合物名の同一性判定へのアプローチ

表記ゆれについては、ある程度パターン化することで回避できる。例えば、「テレフタレート」を「テレフタラート」に統合するといったことである。

表記方法については、化合物名の他に、分子式、示性式、SMILES などの表記法を用いる方法もあるが、一意に決まらなかったり、可読性が低かったりといった問題がある。

日本語の化合物名から化学式を生成する手法についても研究を進めている[9]。また英語の化合物名からSMILES という化合物の表記法を生成する手法についても研究報告がある[10]。

そこで置換基名の違い、命名規則の違いについて、日化辞のデータを利用した言い換えルール辞書を用いて、化合物名の同一性判定をするアプローチを検討した。

4.3　日化辞のデータを利用した言い換えルールの抽出

日化辞には、4.1 節で例を示したように、多くの化合物の体系名や別称が登録されている。日化辞のデータを利用して化合物名のどの部分がどのように言い換えられているのかを調べることで化合物の部分構造の言い換えルールを得ることができる。

例えば、図４のように「アクリル酸 4-tert-ブチルフェニル」と「アクリル酸 4-(1,1-ジメチルエチル）フェニル」の共通部分を除いていくことで、言い換えられている部分「tert-ブチル」と「1,1-ジメチルエチル」を特定することができる。

日化辞全体から 146 万以上の言い換えルールの候補を抽出することができた。ただし、単純な手法によるものなので低出現頻度の言い換えルールは信頼性に乏しい。高出現頻度の言い換えルールについても、「パ」と「ピオ」の言い換えのような意味を持たない文字列の言い換えルールもあったため、文字数が一定以上あるものだけに絞るなどフィルタをかけることで表３のような言い

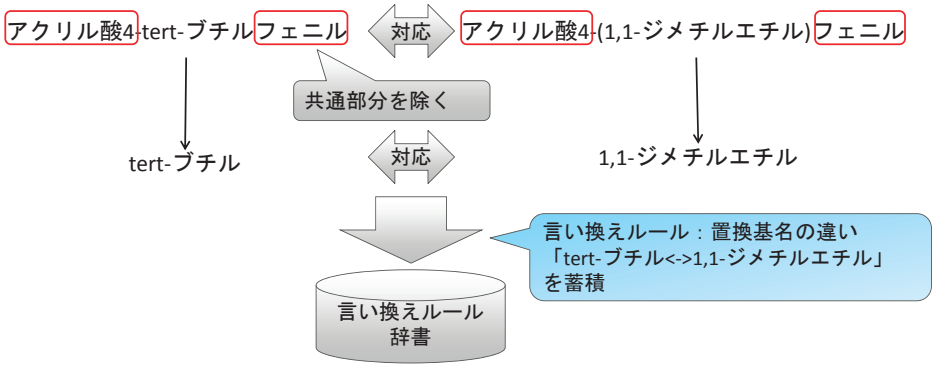

図４　言い換えルールの抽出

表３　高頻度の言い換えルール

出現頻度	言い換えルール：置換基名の違い	備考
2624	プロペン<-> アクリル	異なる構造のため言い換え不可 (「2-プロペン酸」と「アクリル酸」は同じ構造)
1766	エテニル<-> ビニル	同じ構造の別称
1210	ベンゼン<-> フェニル	「ベンゼン」から水素がいくつか抜けたものが「フェニル」
1204	p-トリル<-> 4-メチルフェニル	同じ構造の別称
1176	プロパン酸<-> プロピオン酸	同じ構造の別称
1158	プロペン酸<-> アクリル酸	上位構造として同じ (「2-プロペン酸」と「アクリル酸」は同じ構造)
1126	p-トリル<-> (4-メチルフェニル)	同じ構造の別称
974	アニリン<-> ベンゼンアミン	同じ構造の別称
948	2-プロペニル<-> アリル	同じ構造の別称
947	オクタデカン<-> ステアリン	異なる構造のため言い換え不可 (「オクタデカン酸」と「ステアリン酸」が同じ構造)

換えルールが得られた。上位のものでも信頼性に欠けるものはあるが、妥当なものも多く抽出されている。図４に挙げた「tert-ブチル」と「1,1-ジメチルエチル」の言い換えも 450 回以上出現しており、置換基名の違いをある程度吸収できる可能性がある。

4.4　言い換えルールを用いたクエリ拡張

前節で抽出した言い換えルールを用いて、化合物名を入力として、化合物情報を検索するクエリ拡張システムを試作した。このシステムでは、入力した文字列とカタカナの表記ゆれ、言い換えルールによってできた文字列について、日化辞に対して完全一致で検索を行う。日化辞でヒットすると、日化辞の持つリンク情報から PubChem や ChEMBL などともつながるようになっている。

例えば、このシステムに、「2-(p-トリル）エタノール」を入力した例を図５に示す。この例では、入力された文字列とカタカナの表記ゆれでは日化辞にはヒットせず、言い換えルールによって得られた「2-(4-メチルフェニル）エタノール」では日化辞、PubChem のいずれもヒットした。こうして得られたリンクをたどることで、この化合物の構造や分子量などの情報を得ることができる。

このように化合物 DB にも登録されていない化合物名

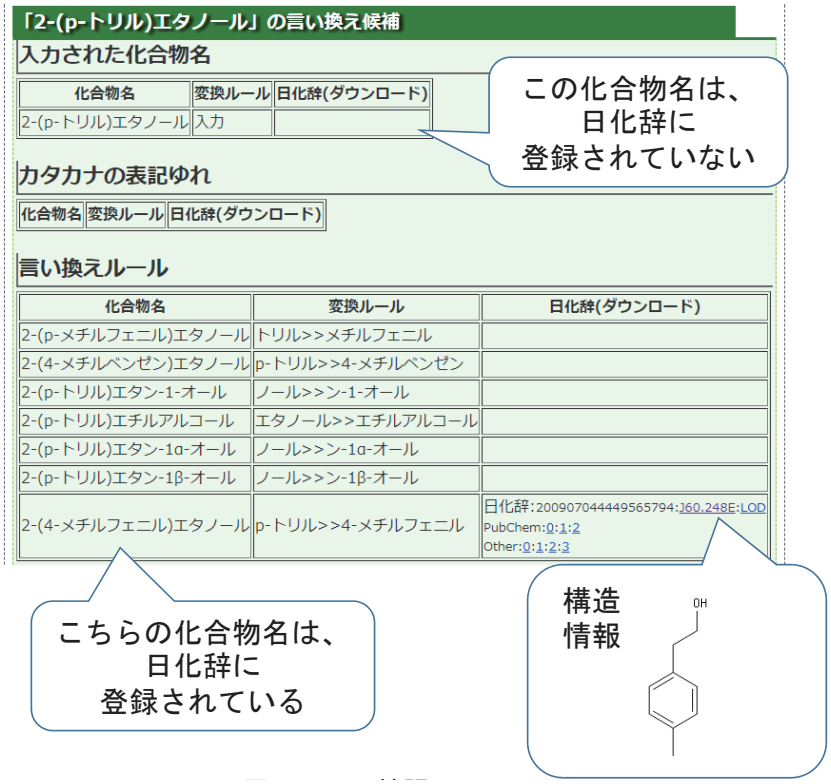

図５　クエリ拡張システム

であっても、言い換えルールを活用することで、目的の化合物のデータへアクセスすることが可能になる。データベースを検索するような場合には、言い換えルールの信頼性が多少低くても必要な情報へアクセスする（再現率を向上させる）ために役立つ。これは、特許を検索する場合にも応用でき、化合物名を入力として特許の検索を行う場合に、このような手法でクエリの拡張を行ってから検索を行うことで、これまでには見つからなかった特許がヒットする可能性がある。より実用的には、言い換えルールを頻度で足きりしたり、言い換え後の化合物名を人手でチェックしたりしたのち検索を行うといった方法で、検索の精度を高めると良いと考えている。言い換えルールの精度が高まれば、特許文書中から抽出した化合物名と化合物 DB とを自動でリンク付けすることもできるようになると考えている。

5 まとめ

化学特許情報や化合物 DB から化合物についての様々な情報を効率よく収集するための基礎技術として、テキストデータから化合物名を抽出する手法を実装し、抽出した化合物名の同一性判定ために化合物名の言い換えルールを構築することにより、化合物の情報をリンクするためのシステムを試作した。日化辞データを利用して構築した言い換えルールにより、化合物名の検索性能向上の可能性を示した。

化合物の同一性判定を可能にすることで、特許、論文、化合物 DB、社内での実験データなど様々な情報をリンクして、より一層の情報活用を実現したい。化合物の情報を集約することができ、研究や調査といった業務に必要な情報に簡単にアクセスできるようになるものと期待する。

今後、大量のテキストデータから化合物と物性値や製造プロセスなどとの関係情報を抽出するとともに、化合物についての知識をナレッジグラフ（グラフ形式で表現された知識ベースであり、実世界における知識を表現、統合、解析することが可能）として連携させることで、利活用の拡大を図りたい。

引用文献

1）内閣府 第３期科学技術基本計画（平成 18～22 年度）
http://www8.cao.go.jp/cstp/kihonkeikaku/kihon3.html
2）IUPAC 命名法
https://iupac.org/what-we-do/nomenclature/
3）日本化学物質辞書（日化辞）
http://dbarchive.biosciencedbc.jp/jp/nikkaji/desc.html
4）ChEMBL　https://www.ebi.ac.uk/chembl/
5）CAS　http://www.cas-japan.jp/
6）The PubChem Project
https://pubchem.ncbi.nlm.nih.gov/
7）ChemSpider　http://www.chemspider.com/
8）Chemical and Product Categories（CPCat）
https://www.epa.gov/chemical-research/chemical-and-product-categories-cpcat
9）池田紀子，田中一成：許文書からの化学物質情報の抽出，Japio YEAR BOOK 2015，p.274-281（2015）
10）SMILES（TM），Simplified Molecular Input Line Entry System
http://www.daylight.com/smiles/index.html